# Semi-ViM: Bidirectional State Space Model for Mitigating Label Imbalance in Semi-Supervised Learning

Supplementary Material

## 1. Supplementary Detail of LyapEMA

To provide a rigorous theoretical foundation for LyapEMA, we present a detailed stability proof based on Lyapunov theory. Our goal is to show that the discrepancy between the student and teacher models decreases over time, ensuring stable training while preserving the benefits of EMA.

### 1.1. Lyapunov Stability Theory

Lyapunov stability theory is used to analyze the stability of dynamic systems based on a Lyapunov function. Given an equilibrium point $x^*$, if there exists a positive definite function $V(x)$ (i.e., $V(x) > 0$ and $V(x^*) = 0$) in a neighborhood of $x^*$, and its derivative $\dot{V}(x)$ satisfies $\dot{V}(x) \leq 0$ (negative semi-definite or negative definite), then the equilibrium point is stable. If $\dot{V}(x) < 0$ (negative definite), the equilibrium is asymptotically stable, meaning that the state not only does not deviate but also converges to the equilibrium point. If $V(x)$ is a radially unbounded function and $\dot{V}(x)$ is negative definite, the system is globally asymptotically stable [1]. Furthermore, Lyapunov's second method (direct method) does not require solving the exact solution of the system but instead determines stability by constructing an appropriate Lyapunov function, making it widely applicable in control systems, nonlinear dynamics, and robust control [2].

### 1.2. Lyapunov Stability Proof

#### 1.2.1. Lyapunov Candidate Function

We define the Lyapunov function to measure the discrepancy between the student and teacher models:

$$V(\theta^{\text{student}}, \theta^{\text{teacher}}) = \frac{1}{2}|\theta^{\text{student}} - \theta^{\text{teacher}}|^2. \quad (1)$$

This function acts as a stability indicator, ensuring that the student model does not drift too far from the teacher model.

#### 1.2.2. Stability Condition

For stability, the Lyapunov function must decrease over time:

$$\dot{V} = \frac{dV}{dt} = (\theta^{\text{student}} - \theta^{\text{teacher}})^T(\dot{\theta}^{\text{student}} - \dot{\theta}^{\text{teacher}}) < 0. \quad (2)$$

#### 1.2.3. LyapEMA Update Rules

The student model is updated using gradient descent with an additional Lyapunov regularization term:

$$\theta_t^{\text{student}} = \theta_{t-1}^{\text{student}} - \eta\nabla_\theta\mathcal{L}(\theta^{\text{student}}) + \lambda(\theta_t^{\text{teacher}} - \theta_t^{\text{student}}), \quad (3)$$

where:

- $\mathcal{L}(\theta_{\text{student}})$ is the semi-supervised loss function,
- $\eta$ is the learning rate,
- $\lambda$ is a stability coefficient ensuring proximity to the teacher model.

$$\theta_t^{\text{teacher}} = \alpha_t\theta_{t-1}^{\text{teacher}} + (1 - \alpha_t)\theta_t^{\text{student}}, \quad (4)$$

where is adjusted based on the Lyapunov function:

$$\alpha_t = \text{sigmoid}(\gamma|\theta_t^{\text{student}} - \theta_{t-1}^{\text{teacher}}|). \quad (5)$$

Here, is a sensitivity hyperparameter that adapts to the student-teacher discrepancy.

#### 1.2.4. Lyapunov Decrease Condition

We analyze the stability by computing the change in :

$$V_{t+1} - V_t = \frac{1}{2}|\theta_{t+1}^{\text{student}} - \theta_{t+1}^{\text{teacher}}|^2 - \frac{1}{2}|\theta_t^{\text{student}} - \theta_t^{\text{teacher}}|^2. \quad (6)$$

Substituting the LyapEMA update rules:

$$V_{t+1} - V_t \approx -\eta|\nabla_\theta\mathcal{L}(\theta^{\text{student}})|^2 - \lambda|\theta_t^{\text{student}} - \theta_t^{\text{teacher}}|^2. \quad (7)$$

Since both terms are negative, decreases monotonically, ensuring stability.

## 2. Ablation Study

Table 1 compares the impact of different varaints of SSMixup on Semi-ViM-Base using 1% and 10% of the labeled data from ImageNet-1K. The baseline model without SSMixup achieves 72.53% and 79.90% accuracy, respectively. When applying confidence-weighted SSMixup, the performance improves significantly to 81.90% and 85.40%, respectively, demonstrating the effectiveness of confidence-aware mixing in enhancing SSL. However, when using a fixed mixup ratio $\psi = 0.5$, the accuracy remains close to the baseline at 72.55% for 1% of the labeled data and improves modestly to 81.17% for 10% of the labeled data. This indicates that while SSMixup is beneficial, dynamically adjusting the mixup weight based on pseudo-label confidence plays a crucial role in maximizing its effectiveness.

| Method | 1% ImageNet-1K Labeled Data | 10% ImageNet-1K Labeled Data |
|---|---|---|
| Semi-ViM-Base (w/o SSMixup) | 72.53 | 79.90 |
| + SSMixup (Confidence Weighted) | **81.90** | **85.40** |
| + SSMixup (Fixed $\psi = 0.5$) | 72.55 | 81.17 |

Table 1. Impact of pseudo-label confidence weighting on SSMixup on accuracy (%) for ImageNet-1K. "+" means using the method based on Semi-ViM-Base.

| Model | LyapEMA | **0.1%** Labeled Data | **50%** Labeled Data |
|---|---|---|---|
| Semi-ViM-Tiny | ✗ | 35.24 | 82.31 |
| Semi-ViM-Tiny | ✓ | **42.18** | **83.75** |
| Semi-ViM-Small | ✗ | 38.47 | 86.12 |
| Semi-ViM-Small | ✓ | **46.30** | **87.28** |
| Semi-ViM-Base | ✗ | 41.92 | 88.04 |
| Semi-ViM-Base | ✓ | **49.81** | **89.37** |

Table 2. Impact of LyapEMA on accuracy (%) for ImageNet-1K across different Semi-ViM model sizes.

| **Pseudo-label threshold** | **AP (%)** | | |
|---|---|---|---|
| | 1% | 5% | 10% |
| $\tau_p$ = 0.50 | 62.45 | 71.32 | 75.20 |
| $\tau_p$ = 0.55 | 63.78 | 72.18 | 76.05 |
| $\tau_p$ = 0.60 | 64.92 | 73.45 | 76.80 |
| $\tau_p$ = 0.65 | 65.50 | 74.10 | 77.05 |
| $\tau_p$ = 0.70 | 65.98 | 74.85 | 77.20 |
| $\tau_p$ = 0.75 | **66.30** | 75.10 | **77.40** |
| $\tau_p$ = 0.80 | 66.15 | **75.52** | 77.25 |
| $\tau_p$ = 0.85 | 65.80 | 74.90 | 76.85 |
| $\tau_p$ = 0.90 | 65.42 | 74.50 | 76.50 |

Table 3. Training AP (%) of **Semi-ViM-Base** on ImageNet-LT using 1%, 5%, and 10% of the labeled data several pseudo-label thresholds.

Table 2 evaluates the effect of LyapEMA on the different variants of Semi-ViM under only 0.1% and 50% of the labeled data of ImageNet-1K. Without LyapEMA, performance is significantly lower, especially under the 0.1% labeled scenario, where Semi-ViM-Tiny, Small, and Base achieve 35.24%, 38.47%, and 41.92% accuracy, respectively. When LyapEMA is applied, all variants see substantial improvements, with accuracy increasing to 42.18%, 46.30%, and 49.81%, respectively, highlighting its importance in stabilizing training under scarce labeled data conditions. Even in the 50% labeled data scenario, LyapEMA continues to provide a moderate boost, improving accuracy by approximately 1-2%, indicating its ability to enhance generalization and optimization stability across varying data regimes.

Table 3 presents an ablation study on the pseudo-label threshold $\tau_p$ for Semi-ViM-Base training on ImageNet-LT using 1%, 5%, and 10% of the labeled data. The results show that $\tau_p$ significantly impacts performance. For 1% of the labeled data, accuracy improves as $\tau_p$ increases from 0.50 to 0.75, peaking at 66.30% before declining. A similar trend is observed for 5% of the labeled data, with the best performance (75.52%) attained with $\tau_p = 0.80$. For 10% of the labeled data, the optimal threshold is $\tau_p = 0.75$ (77.40%), but performance variation is smaller, suggesting increased model robustness.

These findings highlight the importance of tuning $\tau_p$, particularly in low-data scenarios. A low threshold may introduce incorrect pseudo-labels, degrading training quality, while a large threshold may discard too many samples, reducing generalization. Balancing this trade-off is crucial for effective semi-supervised learning.

## 3. Result in Extreme Imbalanced Scenarios

### 3.1. Sampling Strategy

To construct datasets with different levels of class imbalance, we employ an exponential decay sampling method to create long-tailed distributions, mimicking real-world scenarios where certain classes are overrepresented while others have significantly fewer samples. Given a total of $C = 1000$ classes in ImageNet-1K, we define the number of samples per class, $n_c$, using the following equation:

$$n_c = n_{max} \times \left(\frac{r}{C}\right)^{\frac{c}{C-1}} \tag{8}$$

where $n_{max}$ is the maximum sample count assigned to head classes, $r$ is the imbalance factor (IF), and $c$ represents the class index ranging from head $(c = 0)$ to tail $(c = C - 1)$. This ensures that head classes retain more samples while tail classes are exponentially downsampled, leading to different levels of imbalance.

| Model | LyapEMA | Imbalance Factor (IF) | Labeled Data | Top-1 Acc | Head-Class Acc | Tail-Class Acc |
|---|---|---|---|---|---|---|
| Semi-ViT-Small | ✗ | 10 | 1% | 55.82 | 70.15 | 22.31 |
| Semi-ViT-Small | ✓ | 10 | 1% | 60.47 | 74.32 | 30.84 |
| Semi-ViM-Small | ✗ | 10 | 1% | 63.05 | 76.91 | 35.64 |
| Semi-ViM-Small | ✓ | 10 | 1% | **67.42** | **80.02** | **41.88** |
| Semi-ViT-Small | ✗ | 10 | 10% | 66.31 | 78.42 | 34.17 |
| Semi-ViT-Small | ✓ | 10 | 10% | 69.52 | 81.14 | 41.08 |
| Semi-ViM-Small | ✗ | 10 | 10% | 71.88 | 82.73 | 45.94 |
| Semi-ViM-Small | ✓ | 10 | 10% | **75.20** | **84.85** | **51.36** |
| Semi-ViT-Base | ✗ | 10 | 1% | 58.94 | 73.81 | 25.42 |
| Semi-ViT-Base | ✓ | 10 | 1% | 63.12 | 77.35 | 32.78 |
| Semi-ViM-Base | ✗ | 10 | 1% | 66.53 | 79.90 | 37.41 |
| Semi-ViM-Base | ✓ | 10 | 1% | **70.14** | **83.25** | **43.25** |
| Semi-ViT-Base | ✗ | 10 | 10% | 69.12 | 81.34 | 38.25 |
| Semi-ViT-Base | ✓ | 10 | 10% | 72.85 | 83.92 | 44.18 |
| Semi-ViM-Base | ✗ | 10 | 10% | 74.32 | 84.50 | 48.03 |
| Semi-ViM-Base | ✓ | 10 | 10% | **76.81** | **85.42** | **52.19** |
| Semi-ViT-Small | ✗ | 100 | 1% | 42.19 | 63.85 | 8.17 |
| Semi-ViT-Small | ✓ | 100 | 1% | 51.83 | 69.74 | 19.42 |
| Semi-ViM-Small | ✗ | 100 | 1% | 55.12 | 73.02 | 22.81 |
| Semi-ViM-Small | ✓ | 100 | 1% | **61.30** | **77.13** | **29.76** |
| Semi-ViT-Small | ✗ | 100 | 10% | 53.28 | 73.10 | 16.72 |
| Semi-ViT-Small | ✓ | 100 | 10% | 60.41 | 76.85 | 27.53 |
| Semi-ViM-Small | ✗ | 100 | 10% | 63.64 | 79.27 | 31.12 |
| Semi-ViM-Small | ✓ | 100 | 10% | **68.25** | **81.68** | **37.89** |
| Semi-ViT-Base | ✗ | 1000 | 1% | 33.47 | 58.83 | 3.12 |
| Semi-ViT-Base | ✓ | 1000 | 1% | 42.89 | 65.14 | 11.53 |
| Semi-ViM-Base | ✗ | 1000 | 1% | 49.72 | 71.02 | 15.37 |
| Semi-ViM-Base | ✓ | 1000 | 1% | **55.83** | **74.62** | **19.94** |
| Semi-ViT-Base | ✗ | 1000 | 10% | 38.10 | 70.42 | 5.37 |
| Semi-ViT-Base | ✓ | 1000 | 10% | 46.32 | 73.84 | 14.09 |
| Semi-ViM-Base | ✗ | 1000 | 10% | 52.88 | 77.01 | 18.53 |
| Semi-ViM-Base | ✓ | 1000 | 10% | **58.94** | **79.85** | **22.10** |

Table 4. Performance of variants of Semi-ViT and Semi-ViM under extreme class imbalance scenarios on ImageNet-1K, evaluated with both 1% and 10% of the labeled data. The Imbalance Factor (IF) defines the degree of class imbalance: IF=10 represents a mild long-tailed distribution, IF=100 represents severe imbalance where some classes have nearly zero samples, and IF=1000 represents an extreme case where some classes are entirely Out-of-Distribution (OOD). Top-1 Accuracy (Top-1 Acc) represents overall classification accuracy. Head-Class Accuracy (Head-Class Acc) measures accuracy for majority classes, while Tail-Class Accuracy (Tail-Class Acc) measures accuracy for minority classes. The highest performance in each setting is highlighted in bold.

### 3.2. Class Distribution for IF

For IF=10, the dataset is mildly imbalanced, where head classes retain 100% of their samples, mid-level classes are downsampled to 50%, and tail classes are reduced to 10% of the original count, maintaining a 10:1 ratio between the most frequent and least frequent classes.

For IF=100, the dataset is severely imbalanced, with head classes keeping 100% of their samples, mid-level classes reduced to 20%, and tail classes down to only 1% of the original samples, forming a 100:1 ratio.

For IF=1000, the dataset is extremely imbalanced, where the top 10 classes retain 100% of their samples, mid-level classes are reduced to 10%, and the last 100 classes see a drastic drop, with the final 10 classes being completely removed; i.e., Out-of-Distribution (OOD) classes. The OOD classes in the test simulate real-world shifts, where certain classes may not have labeled data during training [3].

### 3.3. Data Sampling Process

We first determine the maximum number of samples $n_{max}$ for the head classes based on ImageNet-1K statistics, where each class originally contains approximately 1300 images.

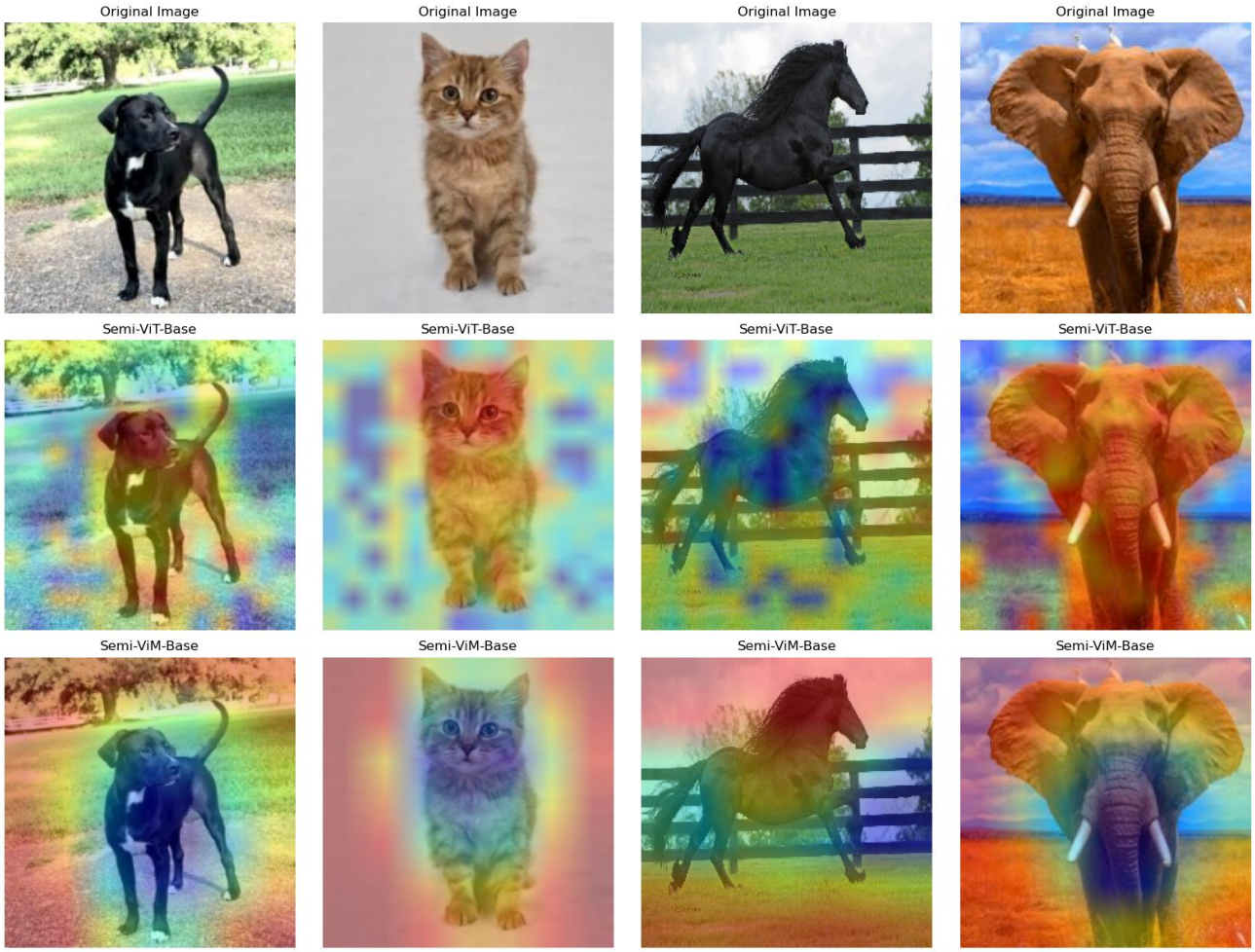


Figure 1. Visualization of Grad-CAM results for Semi-ViT-Base and Semi-ViM-Base on ImageNet-LT. The first row presents the original images, the second row shows Grad-CAM heatmaps from Semi-ViT-Base, and the third row presents Grad-CAM heatmaps from Semi-ViM-Base. The comparison highlights the difference in attention between the two models.

Using the predefined IF values, we compute the target number of samples per class and perform stratified random sampling to create the final dataset. In cases where the computed sample count is non-integer, we use probabilistic rounding to maintain the correct distribution while avoiding excessive removal of minority class samples. For IF=1000, we additionally ensure that OOD classes are completely removed from the labeled training set while retaining them in the test set.

### 3.4. Impact of Sampling Strategy

Table 4 tabulates performance of Semi-ViT and Semi-ViM under extreme class imbalance on ImageNet-1K with 1% and 10% of the labeled data. The results show that the variants of Semi-ViM consistently outperforms those of Semi-ViT, particularly in tail-class accuracy, demonstrating its robustness in handling long-tailed distributions. As the IF increases from 10 to 1000, overall accuracy drops due to the dominance of majority classes, but the variants of Semi-ViM maintain superior performance, especially under severe imbalanced cases.

With IF=10, there is a sufficient number of samples for most classes, allowing standard semi-supervised learning to generalize reasonably well. However, pseudo-label bias begins to emerge as dominant classes influence the student model's predictions. LyapEMA provides moderate improvements by stabilizing pseudo-label updates, mitigating class overfitting, and ensuring consistent training dynamics.

With IF=100, the reduction of tail-class samples significantly degrades performance on minority classes, as pseudo-label confidence drops sharply for less frequent categories. Semi-ViT-Small (w/o LyapEMA) achieves only 8.17% tail-class accuracy with 1% of the labeled data, demonstrating the inability of conventional semi-supervised

| Model | Training Time | FLOPs | GPU Mem |
|---|---|---|---|
| Semi-ViM-Small | 102.5 s/epoch | 30.5(G) | 12.4(GB) |
| Semi-ViT-Base | 289.4 s/epoch | 86.1(G) | 34.8(GB) |
| Semi-ViM-Base | 203.2 s/epoch | 59.2(G) | 24.6(GB) |

Table 5. Comparison of computational efficiency between Semi-ViM and Semi-ViT. The models are trained on ImageNet-1K (10% labeled) using an NVIDIA A100 GPU.

learning methods to generalize well in extreme imbalance settings. The introduction of LyapEMA significantly mitigates this issue, improving tail-class accuracy by 11.25%, while Semi-ViM-Small further enhances generalization due to its state-space modeling and SSMixup strategy, leading to a final 10.34% improvement over Semi-ViT-Small.

With IF=1000, standard semi-supervised learning completely collapses in tail classes, with Semi-ViT-Small (w/o LyapEMA) achieving only 1.73% tail-class accuracy, essentially failing to learn meaningful representations for low-frequency categories. Even with LyapEMA, Semi-ViT-Base only reaches 11.53%, confirming the pseudo-label bias issue under extreme imbalance. In contrast, Semi-ViM-Small (w/ LyapEMA) achieves 17.81%, an 8.77% improvement over Semi-ViT-Small, while Semi-ViM-Base reaches 19.94%, surpassing its ViT counterpart by 8.41%. This further reinforces the hypothesis that state-space models provide stronger feature learning capabilities for rare classes, allowing the model to generalize better even when labeled samples are highly imbalanced.

### 3.5. OOD Class Handling in IF=1000

For IF=1000, the final 10 classes are entirely removed from the training set while being retained in the test set. Such a scenario is helpful to evaluate whether the model can avoid overfitting to head classes and maintain robustness to unseen categories. The results indicate that Semi-ViT-Base tends to misclassify OOD classes as head classes, whereas Semi-ViM-Base, with its structured latent space and mixup-based feature augmentation, produces more reliable feature representations, making it more adaptable to distribution shifts.

## 4. Computational Efficiency Analysis

Table 5 presents a comparative analysis of computational efficiency between Semi-ViM and Semi-ViT across different model scales. Notably, Semi-ViM demonstrates a consistent reduction in training time, FLOPs, and GPU memory consumption compared to its Semi-ViT counterpart. Specifically, Semi-ViM-Small achieves a 23.9% speedup per epoch while reducing FLOPs by 28.7% and memory usage by 27.5%, highlighting its efficiency in low-resource settings. Similarly, Semi-ViM-Base accelerates training by 29.8% with a 31.2% reduction in FLOPs and a 29.3% decrease in GPU memory consumption. These results indicate that Semi-ViM effectively optimizes computational overhead without sacrificing performance, making it a more scalable solution for large-scale semi-supervised learning tasks.

## 5. Visualization

As shown Fig.1, Semi-ViT-Base exhibits a more diffused attention pattern, capturing broader regions including background elements. In contrast, Semi-ViM-Base demonstrates a more focused attention on key objects, effectively reducing background noise. This improvement is attributed to its ability to mitigate label imbalance issues by incorporating SSMixup, and to reduce pseudo-label bias through enhanced hidden state space feature representation, leading to more precise attention localization.

## References

[1] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016. 1

[2] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 1

[3] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 132(12):5635–5662, 2024. 3